10-28-04

A 17ew

PTO/SB/21 (04-04)

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	10/777,832
Filing Date	February 11, 2004
First Named Inventor	YAGISAWA, Ikuya
Art Unit	2818
Examiner Name	Unassigned
Total Number of Pages in This Submission: 9	Attorney Docket Number: 16869P-105400US

ENCLOSURES *(Check all that apply)*

- ☒ Fee Transmittal Form
  - ☐ Fee Attached
- ☐ Amendment/Reply
  - ☐ After Final
  - ☐ Affidavits/declaration(s)
- ☐ Extension of Time Request
- ☐ Express Abandonment Request
- ☐ Information Disclosure Statement
- ☐ Certified Copy of Priority Document(s)
- ☐ Response to Missing Parts/ Incomplete Application
  - ☐ Response to Missing Parts under 37 CFR 1.52 or 1.53
- ☐ Drawing(s)
- ☐ Licensing-related Papers
- ☒ Petition to Make Special
- ☐ Petition to Convert to a Provisional Application
- ☐ Power of Attorney, Revocation Change of Correspondence Address
- ☐ Terminal Disclaimer
- ☐ Request for Refund
- ☐ CD, Number of CD(s) _____
- ☐ After Allowance Communication to Technology Center (TC)
- ☐ Appeal Communication to Board of Appeals and Interferences
- ☐ Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- ☐ Proprietary Information
- ☐ Status Letter
- ☒ Other Enclosure(s) (please identify below):
  - Return Postcard
  - Five (5) cited references

Remarks The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual name	Townsend and Townsend and Crew LLP Chun-Pok Leung Reg. No. 41,405
Signature	
Date	October 26, 2004

CERTIFICATE OF TRANSMISSION/MAILING

Express Mail Label: **EV 530886980 US**

I hereby certify that this correspondence is being deposited with the United States Postal Service with "Express Mail Post Office to Address" service under 37 CFR 1.10 on this date **October 26, 2004** and is addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.

Typed or printed name	Joy Salvador		
Signature		Date	October 26, 2004

60340797 v1

FEE TRANSMITTAL for FY 2005

Effective 10/01/2004. Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 130.00

Complete if Known	
Application Number	10/777,832
Filing Date	February 11, 2004
First Named Inventor	YAGISAWA, Ikuya
Examiner Name	Unassigned
Art Unit	2818
Attorney Docket No.	16869P-105400US

METHOD OF PAYMENT *(check all that apply)*

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:

Deposit Account Number: 20-1430

Deposit Account Name: Townsend and Townsend and Crew LLP

The Director is authorized to: *(check all that apply)*

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, **except for the filing fee** to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1001	790	2001	395	Utility filing fee	
1002	350	2002	175	Design filing fee	
1003	550	2003	275	Plant filing fee	
1004	790	2004	395	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	

SUBTOTAL (1) ($)0.00

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Extra Claims		Fee from below		Fee Paid
Total Claims		-** =		X		=	
Independent Claims		-** =		X		=	
Multiple Dependent				X		=	

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
1202	18	2202	9	Claims in excess of 20
1201	88	2201	44	Independent claims in excess of 3
1203	300	2203	150	Multiple dependent claim, if not paid
1204	88	2204	44	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)0.00

**or number previously paid, if greater; For Reissues, see above

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet.	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for *ex parte* reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	430	2252	215	Extension for reply within second month	
1253	980	2253	490	Extension for reply within third month	
1254	1,530	2254	765	Extension for reply within fourth month	
1255	2,080	2255	1,040	Extension for reply within fifth month	
1401	340	2401	170	Notice of Appeal	
1402	340	2402	170	Filing a brief in support of an appeal	
1403	300	2403	150	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive – unavoidable	
1453	1,330	2453	665	Petition to revive – unintentional	
1501	1,370	2501	685	Utility issue fee (or reissue)	
1502	490	2502	245	Design issue fee	
1503	660	2503	330	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	130
1807	50	1807	50	Processing fee under 37 CFR 1.17(q)	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	790	2809	395	Filing a submission after final rejection (37 CFR § 1.129(a))	
1810	790	2810	395	For each additional invention to be examined (37 CFR § 1.129(b))	
1801	790	2801	395	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid **SUBTOTAL (3)** ($)130.00

SUBMITTED BY				Complete *(if applicable)*	
Name (Print/Type)	Chun-Pok Leung	*Registration No. (Attorney/Agent)*	41,405	*Telephone*	650-326-2400
Signature				*Date*	October 26, 2004

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

60340290 v1

PATENT
Attorney Docket No.: 16869P-105400US
Client Ref. No.: 340301085US011

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

IKUYA YAGISAWA *et al.*

Application No.: 10/777,832

Filed: February 11, 2004

For: DISK ARRAY OPTIMIZING THE DRIVE OPERATION TIME

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2818

Confirmation No.: 5908

PETITION TO MAKE SPECIAL FOR NEW APPLICATION UNDER M.P.E.P. § 708.02, VIII & 37 C.F.R. § 1.102(d)

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of $130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

11/01/2004 BABRAHA1 00000005 201430 10777832
01 FC:1460 130.00 DA

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicants will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search, a computer database search, a keyword search, a literature search, and a foreign patent document search. The searches were performed on or around August 19, 2004, and were conducted by a professional search firm, Kramer & Amado, P.C. The classification search covered Classes 711 (subclass 113), 713 (subclasses 310 and 321), and 714 (subclass 6). The computer database search was conducted on the USPTO systems EAST and WEST. The keyword search was conducted in Classes 710 (subclass 5); 711 (subclasses 112, 114, 154, and 162); 713 (subclasses 300, 320, 322, and 323), and 714 (subclasses 5 and 7). The literature search was conducted on the Internet. The search for foreign patent documents was conducted on the Espacenet and Delphion databases. The inventors further provided two references considered most closely related to the subject matter of the present application (see references #4 and #5 below), which were cited in the Information Disclosure Statement filed with the application on February 11, 2004.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

(1) U.S. Patent No. 5,900,007;
(2) U.S. Patent No. 5,461,266;
(3) U.S. Patent No. 5,734,912;
(4) David A. Patterson et al., "A Case for Redundant Arrays of Inexpensive Disks (RAID); and
(5) Japanese Patent Publication No. 2000-293314.

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to an external storage device system and, more particularly, to a technology for prolonging an operation period of a disk device (hereafter also referred to simply as a disk) and decreasing power consumption of a storage device system (hereafter referred to as a disk array). The disk device's operation period signifies a period from the time to start using the disk device to the time when the disk device becomes unusable.

Independent claim 1 recites a storage system connected to a computer. The storage system comprises a plurality of logical units comprising disk devices. The storage system receives an instruction from the computer to turn on or off a disk device corresponding to the logical unit. Based on the instruction, the storage system turns on or off the disk device corresponding to the logical unit independently of disk devices corresponding to the other logical units.

Independent claim 7 recites a computer system comprising a computer; and a storage system. The storage system has a plurality of logical units comprising disk devices. The computer provides the storage system with an instruction to turn on or off a disk device corresponding to the logical unit. The storage system receives the instruction; and turns on or off the disk device corresponding to the logical unit based on the instruction independently of disk devices corresponding to the other logical units.

Independent claim 14 recites a computer program product for a computer system comprising a computer and a storage system having a plurality of logical units comprising disk devices. The computer program product comprises code for the computer to provide the storage system with an instruction to turn on or off a disk device corresponding to the logical unit; code for the storage system to receive the instruction; code for the storage system to turn on or off a disk device corresponding to the logical unit based on the instruction independently of disk devices corresponding to the other logical units; and a computer readable storage medium for storing the codes.

One benefit that may be derived is prolonging operation times of disk devices constituting a disk array and decreasing the disk array's power consumption.

B. Discussion of the References

None of the following references disclose or suggest a storage system that receives an instruction from the computer to turn on or off a disk device corresponding to the logical unit and, based on the instruction, turns on or off the disk device corresponding to the logical unit independently of disk devices corresponding to the other logical units.

1. U.S. Patent No. 5,900,007

This reference discloses a data storage and retrieval system that includes a large array of small disk files, and three storage managers for controlling the allocation of data to the array, access to data, and the power status of disk files within the array. The operation of the data storage system centers around power management subsystem 106, which manages disk array 110 such that at any point in time some disk files are active (power-on) and others are inactive (power-off), and further such that the disk files which are active are those determined to be the best suited to serving the read and write storage requests pending in the system at that time. See column 3, line 46 to column 4, line 51.

2. U.S. Patent No. 5,461,266

This reference discloses a system for controlling power consumption for an information processing apparatus. Reducing the power consumed by the floppy disk drive or the hard disk drive is achieved by monitoring the use of the disk drive by means of an exclusive CPU and automatically stopping a motor for the drive if there has been no access thereto for a given period of time.

3. U.S. Patent No. 5,734,912

This reference relates to a power control apparatus for an input/output subsystem comprising an input/output control section, which is provided in a disk unit, and performs control of data input/output to and from the plurality of disk modules in the same unit and issuance, upon power-on, of a power-on instruction in compliance with a predetermined procedure.

4. David A. Patterson et al., "A Case for Redundant Arrays of Inexpensive Disks (RAID)

This reference discloses a disk array as a type of storage device systems connected to a computer. The disk array is also referred to as a RAID (Redundant Arrays of Inexpensive Disks) and constitutes a storage device system comprising a plurality of disk devices arranged in an array and a control section to control them. The disk array concurrently operates disk devices to accelerate read requests (requests to read data) and write requests (requests to write data) and to provide data with redundancy. Disk arrays are categorized into five levels depending on types of redundant data to be added and disk array configurations.

5. Japanese Patent Publication No. 2000-293314

This reference discloses a technique to suppress the power consumption of a magnetic disk drive mounted on a disk array device. The device is provided with a means which controls the relation between the configuration of plural magnetic disk drives and access from a host device 101, a power-saving controlling means which controls the power-saving (selection of power on/off and power-saving mode) of magnetic disk drives in a set logical drive, and a controlling means which controls the diagnoses of the magnetic disk drives. This disk array device 110 shifts a prescribed magnetic disk drive to a power-saving mode or turns off the power (power-saving processing) after access from the device 101 does not exist any more and a predetermined time elapses. The magnetic disk drive undergoing power-save processing is subjected to diagnosis execution after a prescribed time passes at the start of the power-saving processing or when a designated time comes in order to maintain its reliability.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,

Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60307840 v1

A Case for Redundant Arrays of Inexpensive Disks (RAID)

David A Patterson, Garth Gibson, and Randy H Katz

Computer Science Division
Department of Electrical Engineering and Computer Sciences
571 Evans Hall
University of California
Berkeley, CA 94720
(pattrsn@ginger berkeley edu)

Abstract Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability This paper introduces five levels of RAIDs, giving their relative cost/performance, and compares RAID to an IBM 3380 and a Fujitsu Super Eagle

1 Background: Rising CPU and Memory Performance

The users of computers are currently enjoying unprecedented growth in the speed of computers Gordon Bell said that between 1974 and 1984, single chip computers improved in performance by 40% per year, about twice the rate of minicomputers [Bell 84] In the following year Bill Joy predicted an even faster growth [Joy 85]

$$MIPS = 2^{Year-1984}$$

Mainframe and supercomputer manufacturers, having difficulty keeping pace with the rapid growth predicted by "Joy's Law," cope by offering multiprocessors as their top-of-the-line product.

But a fast CPU does not a fast system make Gene Amdahl related CPU speed to main memory size using this rule [Siewiorek 82]

Each CPU instruction per second requires one byte of main memory,

If computer system costs are not to be dominated by the cost of memory, then Amdahl's constant suggests that memory chip capacity should grow at the same rate Gordon Moore predicted that growth rate over 20 years ago

$$transistors/chip = 2^{Year-1964}$$

As predicted by Moore's Law, RAMs have quadrupled in capacity every two [Moore 75] to three years [Myers 86]

Recently the ratio of megabytes of main memory to MIPS has been defined as *alpha* [Garcia 84], with Amdahl's constant meaning alpha = 1 In part because of the rapid drop of memory prices, main memory sizes have grown faster than CPU speeds and many machines are shipped today with alphas of 3 or higher

To maintain the balance of costs in computer systems, secondary storage must match the advances in other parts of the system A key measure of magnetic disk technology is the growth in the maximum number of bits that can be stored per square inch, or the bits per inch in a track times the number of tracks per inch Called M A D , for maximal areal density, the "First Law in Disk Density" predicts [Frank87]

$$MAD = 10^{(Year-1971)/10}$$

Magnetic disk technology has doubled capacity and halved price every three years, in line with the growth rate of semiconductor memory, and in practice between 1967 and 1979 the disk capacity of the average IBM data processing system more than kept up with its main memory [Stevens81]

Capacity is not the only memory characteristic that must grow rapidly to maintain system balance, since the speed with which instructions and data are delivered to a CPU also determines its ultimate performance The speed of main memory has kept pace for two reasons

(1) the invention of caches, showing that a small buffer can be managed automatically to contain a substantial fraction of memory references,
(2) and the SRAM technology, used to build caches, whose speed has improved at the rate of 40% to 100% per year

In contrast to primary memory technologies, the performance of single large expensive magnetic disks (SLED) has improved at a modest rate These *mechanical* devices are dominated by the seek and the rotation delays from 1971 to 1981, the raw seek time for a high-end IBM disk improved by only a factor of two while the rotation time did not change[Harker81] Greater density means a higher transfer rate when the information is found, and extra heads can reduce the average seek time, but the raw seek time only improved at a rate of 7% per year There is no reason to expect a faster rate in the near future

To maintain balance, computer systems have been using even larger main memories or solid state disks to buffer some of the I/O activity This may be a fine solution for applications whose I/O activity has locality of reference and for which volatility is not an issue, but applications dominated by a high rate of random requests for small pieces of data (such as transaction-processing) or by a low number of requests for massive amounts of data (such as large simulations running on supercomputers) are facing a serious performance limitation

2. The Pending I/O Crisis

What is the impact of improving the performance of some pieces of a problem while leaving others the same? Amdahl's answer is now known as Amdahl's Law [Amdahl67]

$$S = \frac{1}{(1-f) + f/k}$$

where

S = the effective speedup,
f = fraction of work in faster mode, and
k = speedup while in faster mode

Suppose that some current applications spend 10% of their time in I/O Then when computers are 10X faster--according to Bill Joy in just over three years--then Amdahl's Law predicts effective speedup will be only 5X When we have computers 100X faster--via evolution of uniprocessors or by multiprocessors--this application will be less than 10X faster, wasting 90% of the potential speedup

While we can imagine improvements in software file systems via buffering for near term I/O demands, we need innovation to avoid an I/O crisis [Boral 83]

3 A Solution: Arrays of Inexpensive Disks

Rapid improvements in capacity of large disks have not been the only target of disk designers, since personal computers have created a market for inexpensive magnetic disks These lower cost disks have lower performance as well as less capacity Table I below compares the top-of-the-line IBM 3380 model AK4 mainframe disk, Fujitsu M2361A "Super Eagle" minicomputer disk, and the Conner Peripherals CP 3100 personal computer disk

Characteristics	*IBM 3380*	*Fujitsu M2361A*	*Conners CP3100*	*3380 v 3100*	*2361 v 3100*
				(>1 means 3100 is better)	
Disk diameter (inches)	14	10 5	3 5	4	3
Formatted Data Capacity (MB)	7500	600	100	01	2
Price/MB(controller incl)	$18-$10	$20-$17	$10-$7	1-2 5	1 7-3
MTTF Rated (hours)	30,000	20,000	30,000	1	1 5
MTTF in practice (hours)	100,000	?	?	?	?
No Actuators	4	1	1	2	1
Maximum I/O's/second/Actuator	50	40	30	6	8
Typical I/O's/second/Actuator	30	24	20	7	8
Maximum I/O's/second/box	200	40	30	2	8
Typical I/O's/second/box	120	24	20	2	8
Transfer Rate (MB/sec)	3	2 5	1	3	4
Power/box (W)	6,600	640	10	660	64
Volume (cu ft)	24	3 4	03	800	110

Table I *Comparison of IBM 3380 disk model AK4 for mainframe computers, the Fujitsu M2361A "Super Eagle" disk for minicomputers, and the Conners Peripherals CP 3100 disk for personal computers By "Maximum I/O's/second" we mean the maximum number of average seeks and average rotates for a single sector access Cost and reliability information on the 3380 comes from widespread experience [IBM 87] [Gawlick87] and the information on the Fujitsu from the manual [Fujitsu 87], while some numbers on the new CP3100 are based on speculation The price per megabyte is given as a range to allow for different prices for volume discount and different mark-up practices of the vendors (The 8 watt maximum power of the CP3100 was increased to 10 watts to allow for the inefficiency of an external power supply, since the other drives contain their own power supplies)*

One surprising fact is that the number of I/Os per second per actuator in an inexpensive disk is within a factor of two of the large disks In several of the remaining metrics, including price per megabyte, the inexpensive disk is superior or equal to the large disks

The small size and low power are even more impressive since disks such as the CP3100 contain full track buffers *and* most functions of the traditional mainframe controller Small disk manufacturers can provide such functions in high volume disks because of the efforts of standards committees in defining higher level peripheral interfaces, such as the ANSI X3 131-1986 Small Computer System Interface (SCSI) Such standards have encouraged companies like Adeptec to offer SCSI interfaces as single chips, in turn allowing disk companies to *embed* mainframe controller functions at low cost Figure 1 compares the traditional mainframe disk approach and the small computer disk approach The same SCSI interface chip embedded as a controller in every disk can also be used as the direct memory access (DMA) device at the other end of the SCSI bus

Such characteristics lead to our proposal for building I/O systems as arrays of inexpensive disks, either interleaved for the large transfers of supercomputers [Kim 86][Livny 87][Salem86] or independent for the many small transfers of transaction processing Using the information in Table I, 75 inexpensive disks potentially have 12 times the I/O bandwidth of the IBM 3380 and the same capacity, with lower power consumption and cost

4 Caveats

We cannot explore all issues associated with such arrays in the space available for this paper, so we concentrate on fundamental estimates of price-performance and reliability Our reasoning is that if there are no advantages in price-performance or terrible disadvantages in reliability, then there is no need to explore further We characterize a transaction-processing workload to evaluate performance of a collection of inexpensive disks, but remember that such a collection is just one hardware component of a complete tranaction-processing system While designing a complete TPS based on these ideas is enticing, we will resist that temptation in this paper Cabling and packaging, certainly an issue in the cost and reliability of an array of many inexpensive disks, is also beyond this paper's scope

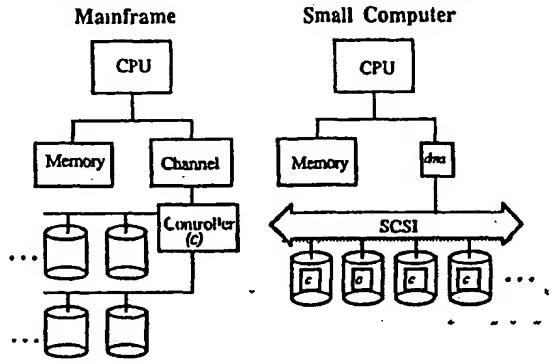


Figure 1 *Comparison of organizations for typical mainframe and small computer disk interfaces Single chip SCSI interfaces such as the Adaptec AIC-6250 allow the small computer to use a single chip to be the DMA interface as well as provide an embedded controller for each disk [Adeptec 87] (The price per megabyte in Table I includes everything in the shaded boxes above)*

5. And Now The Bad News: Reliability

The unreliability of disks forces computer systems managers to make backup versions of information quite frequently in case of failure What would be the impact on reliability of having a hundredfold increase in disks? Assuming a constant failure rate--that is, an exponentially distributed time to failure--and that failures are independent--both assumptions made by disk manufacturers when calculating the Mean Time To Failure (MTTF)--the reliability of an array of disks is

$$\textit{MTTF of a Disk Array} = \frac{\textit{MTTF of a Single Disk}}{\textit{Number of Disks in the Array}}$$

Using the information in Table I, the MTTF of 100 CP 3100 disks is 30,000/100 = 300 hours, or less than 2 weeks Compared to the 30,000 hour (> 3 years) MTTF of the IBM 3380, this is dismal If we consider scaling the array to 1000 disks, then the MTTF is 30 hours or about one day, requiring an adjective worse than dismal

Without fault tolerance, large arrays of inexpensive disks are too unreliable to be useful

6. A Better Solution: RAID

To overcome the reliability challenge, we must make use of extra disks containing redundant information to recover the original information when a disk fails Our acronym for these Redundant Arrays of Inexpensive Disks is ***RAID*** To simplify the explanation of our final proposal and to avoid confusion with previous work, we give a taxonomy of five different organizations of disk arrays, beginning with mirrored disks and progressing through a variety of alternatives with differing performance and reliability We refer to each organization as a RAID *level*

The reader should be forewarned that we describe all levels as if implemented in hardware solely to simplify the presentation, for RAID ideas are applicable to software implementations as well as hardware

Reliability Our basic approach will be to break the arrays into reliability groups, with each group having extra "check" disks containing redundant information When a disk fails we assume that within a short time the failed disk will be replaced and the information will be

reconstructed on to the new disk using the redundant information This time is called the mean time to repair (MTTR) The MTTR can be reduced if the system includes extra disks to act as "hot" standby spares, when a disk fails, a replacement disk is switched in electronically Periodically a human operator replaces all failed disks Here are other terms that we use

D = total number of disks with data (not including extra check disks),
G = number of data disks in a *group* (not including extra check disks),
C = number of check disks in a group,
$n_G = D/G$ = number of groups,

As mentioned above we make the same assumptions that disk manufacturers make--that failures are exponential and independent (An earthquake or power surge is a situation where an array of disks might not fail independently) Since these reliability predictions will be very high, we want to emphasize that the reliability is only of the the disk-head assemblies with this failure model, and not the whole software and electronic system In addition, in our view the pace of technology means extremely high MTTF are "overkill"--for, independent of expected lifetime, users will replace obsolete disks After all, how many people are still using 20 year old disks?

The general MTTF calculation for single-error repairing RAID is given in two steps First, the group MTTF is

$$MTTF_{Group} = \frac{MTTF_{Disk}}{G+C} * \frac{1}{\text{Probability of another failure in a group before repairing the dead disk}}$$

As more formally derived in the appendix, the probability of a second failure before the first has been repaired is

$$\text{Probability of Another Failure} = \frac{MTTR}{MTTF_{Disk}/(\text{No Disks}-1)} = \frac{MTTR}{MTTF_{Disk}/(G+C-1)}$$

The intuition behind the formal calculation in the appendix comes from trying to calculate the average number of second disk failures during the repair time for X single disk failures Since we assume that disk failures occur at a uniform rate, this average number of second failures during the repair time for X first failures is

$$\frac{X * MTTR}{MTTF \text{ of remaining disks in the group}}$$

The average number of second failures for a single disk is then

$$\frac{MTTR}{MTTF_{Disk}/\text{No of remaining disks in the group}}$$

The MTTF of the remaining disks is just the MTTF of a single disk divided by the number of good disks in the group, giving the result above

The second step is the reliability of the whole system, which is approximately (since $MTTF_{Group}$ is not quite distributed exponentially)

$$MTTF_{RAID} = \frac{MTTF_{Group}}{n_G}$$

Plugging it all together, we get

$$MTTF_{RAID} = \frac{MTTF_{Disk}}{G+C} * \frac{MTTF_{Disk}}{(G+C-1)*MTTR} * \frac{1}{n_G}$$

$$= \frac{(MTTF_{Disk})^2}{(G+C)*n_G*(G+C-1)*MTTR}$$

$$MTTF_{RAID} = \frac{(MTTF_{Disk})^2}{(D+C*n_G)*(G+C-1)*MTTR}$$

Since the formula is the same for each level, we make the abstract numbers concrete using these parameters as appropriate D=100 total data disks, G=10 data disks per group, $MTTF_{Disk}$ = 30,000 hours, $MTTR$ = 1 hour, with the check disks per group C determined by the RAID level

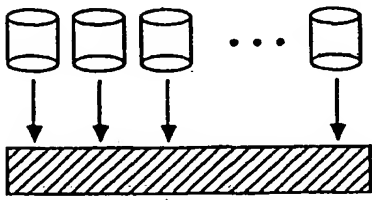
Reliability Overhead Cost This is simply the extra check disks, expressed as a percentage of the number of data disks D As we shall see below, the cost varies with RAID level from 100% down to 4%

Useable Storage Capacity Percentage Another way to express this reliability overhead is in terms of the percentage of the total capacity of data disks *and* check disks that can be used to store data Depending on the organization, this varies from a low of 50% to a high of 96%

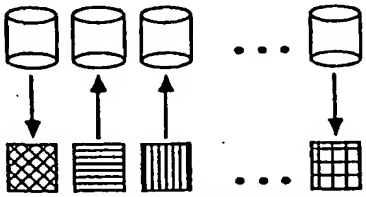
Performance Since supercomputer applications and transaction-processing systems have different access patterns and rates, we need different metrics to evaluate both For supercomputers we count the number of reads and writes per second for large blocks of data, with large defined as getting at least one sector from each data disk in a group During large transfers all the disks in a group act as a single unit, each reading or writing a portion of the large data block in parallel

A better measure for transaction-processing systems is the number of individual reads or writes per second Since transaction-processing systems (e g , debits/credits) use a read-modify-write sequence of disk accesses, we include that metric as well Ideally during small transfers each disk in a group can act independently, either reading or writing independent information In summary supercomputer applications need a *high data rate* while transaction-processing need a *high I/O rate*

For both the large and small transfer calculations we assume the minimum user request is a sector, that a sector is small relative to a track, and that there is enough work to keep every device busy Thus sector size affects both disk storage efficiency and transfer size Figure 2 shows the ideal operation of large and small disk accesses in a RAID



(a) Single Large or "Grouped" Read (1 read spread over G disks)



(b) Several Small or Individual Reads and Writes (G reads and/or writes spread over G disks)

Figure 2. *Large transfer vs small transfers in a group of G disks*

The six performance metrics are then the number of reads, writes, and read-modify-writes per second for both large (grouped) or small (individual) transfers Rather than give absolute numbers for each metric, we calculate efficiency the number of events per second for a RAID relative to the corresponding events per second for a single disk (This is Boral's I/O bandwidth per gigabyte [Boral 83] scaled to gigabytes per disk) In this paper we are after fundamental differences so we use simple, deterministic throughput measures for our performance metric rather than latency

Effective Performance Per Disk The cost of disks can be a large portion of the cost of a database system, so the I/O performance per disk--factoring in the overhead of the check disks--suggests the cost/performance of a system This is the bottom line for a RAID

7. First Level RAID: Mirrored Disks

Mirrored disks are a traditional approach for improving reliability of magnetic disks This is the most expensive option we consider since all disks are duplicated (G=1 and C=1), and every write to a data disk is also a write to a check disk Tandem doubles the number of controllers for fault tolerance, allowing an optimized version of mirrored disks that lets reads occur in parallel Table II shows the metrics for a Level 1 RAID assuming this optimization

MTTF	Exceeds Useful Product Lifetime (4,500,000 hrs or > 500 years)
Total Number of Disks	2D
Overhead Cost	100%
Useable Storage Capacity	50%

Events/Sec vs Single Disk	*Full RAID*	*Efficiency Per Disk*
Large (or Grouped) Reads	*2D/S*	1 00/S
Large (or Grouped) Writes	*D/S*	50/S
Large (or Grouped) R-M-W	*4D/3S*	67/S
Small (or Individual) Reads	*2D*	1 00
Small (or Individual) Writes	*D*	50
Small (or Individual) R-M-W	*4D/3*	67

Table II. *Characteristics of Level 1 RAID Here we assume that writes are not slowed by waiting for the second write to complete because the slowdown for writing 2 disks is minor compared to the slowdown S for writing a whole group of 10 to 25 disks Unlike a "pure" mirrored scheme with extra disks that are invisible to the software, we assume an optimized scheme with twice as many controllers allowing parallel reads to all disks, giving full disk bandwidth for large reads and allowing the reads of read-modify-writes to occur in parallel*

When individual accesses are distributed across multiple disks, average queueing, seek, and rotate delays may differ from the single disk case Although bandwidth may be unchanged, it is distributed more evenly, reducing variance in queueing delay and, if the disk load is not too high, also reducing the expected queueing delay through parallelism [Livny 87] When many arms seek to the same track then rotate to the described sector, the average seek and rotate time will be larger than the average for a single disk, tending toward the worst case times This affect should not generally more than double the average access time to a single sector while still getting many sectors in parallel In the special case of mirrored disks with sufficient controllers, the choice between arms that can read any data sector will reduce the time for the average read seek by up to 45% [Bitton 88]

To allow for these factors but to retain our fundamental emphasis we apply a slowdown factor, S, when there are more than two disks in a group In general, $1 \le S \le 2$ whenever groups of disk work in parallel With synchronous disks the spindles of all disks in the group are synchronous so that the corresponding sectors of a group of disks pass under the heads simultaneously,[Kurzweil 88] so for synchronous disks there is no slowdown and $S = 1$ Since a Level 1 RAID has only one data disk in its group, we assume that the large transfer requires the same number of disks acting in concert as found in groups of the higher level RAIDs 10 to 25 disks

Duplicating all disks can mean doubling the cost of the database system or using only 50% of the disk storage capacity Such largess inspires the next levels of RAID

8 Second Level RAID: Hamming Code for ECC

The history of main memory organizations suggests a way to reduce the cost of reliability With the introduction of 4K and 16K DRAMs, computer designers discovered that these new devices were subject to losing information due to alpha particles Since there were many single bit DRAMs in a system and since they were usually accessed in groups of 16 to 64 chips at a time, system designers added redundant chips to correct single errors and to detect double errors in each group This increased the number of memory chips by 12% to 38%--depending on the size of the group--but it significantly improved reliability

As long as all the data bits in a group are read or written together, there is no impact on performance However, reads of less than the group size require reading the whole group to be sure the information is correct, and writes to a portion of the group mean three steps

1) a read step to get all the rest of the data,
2) a modify step to merge the new and old information,
3) a write step to write the full group, including check information

Since we have scores of disks in a RAID and since some accesses are to groups of disks, we can mimic the DRAM solution by bit-interleaving the data across the disks of a group and then add enough check disks to detect and correct a single error A single parity disk can detect a single error, but to correct an error we need enough check disks to identify the disk with the error For a group size of 10 data disks (G) we need 4 check disks (C) in total, and if $G = 25$ then $C = 5$ [Hamming50] To keep down the cost of redundancy, we assume the group size will vary from 10 to 25

Since our individual data transfer unit is just a sector, bit- interleaved disks mean that a large transfer for this RAID must be at least G sectors Like DRAMs, reads to a smaller amount implies reading a full sector from each of the bit-interleaved disks in a group, and writes of a single unit involve the read-modify-write cycle to all the disks Table III shows the metrics of this Level 2 RAID

MTTF		Exceeds Useful Lifetime			
		G=10 (494,500 hrs or >50 years)		*G=25* (103,500 hrs or 12 years)	
Total Number of Disks		1 40D		1.20D	
Overhead Cost		40%		20%	
Useable Storage Capacity		71%		83%	
Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk* *L2*	*L2/L1*	*Efficiency Per Disk* *L2*	*L2/L1*
Large Reads	*D/S*	71/S	71%	86/S	86%
Large Writes	*D/S*	71/S	143%	86/S	172%
Large R-M-W	*D/S*	71/S	107%	86/S	129%
Small Reads	*D/SG*	07/S	6%	03/S	3%
Small Writes	*D/2SG*	04/S	6%	02/S	3%
Small R-M-W	*D/SG*	07/S	9%	03/S	4%

Table III *Characteristics of a Level 2 RAID The L2/L1 column gives the % performance of level 2 in terms of level 1 (>100% means L2 is faster) As long as the transfer unit is large enough to spread over all the data disks of a group, the large I/Os get the full bandwidth of each disk, divided by S to allow all disks in a group to complete Level 1 large reads are faster because data is duplicated and so the redundancy disks can also do independent accesses Small I/Os still require accessing all the disks in a group, so only D/G small I/Os can happen at a time, again divided by S to allow a group of disks to finish Small Level 2 writes are like small R-M-W because full sectors must be read before new data can be written onto part of each sector*

For large writes, the level 2 system has the same performance as level 1 even though it uses fewer check disks, and so on a per disk basis it outperforms level 1 For small data transfers the performance is dismal either for the whole system or per disk, all the disks of a group must be accessed for a small transfer, limiting the maximum number of simultaneous accesses to D/G We also include the slowdown factor S since the access must wait for all the disks to complete

Thus level 2 RAID is desirable for supercomputers but inappropriate for transaction processing systems, with increasing group size increasing the disparity in performance per disk for the two applications In recognition of this fact, Thinking Machines Incorporated announced a Level 2 RAID this year for its Connection Machine supercomputer called the "Data Vault," with G = 32 and C = 8, including one hot standby spare [Hillis 87]

Before improving small data transfers, we concentrate once more on lowering the cost

9 Third Level RAID: Single Check Disk Per Group

Most check disks in the level 2 RAID are used to determine which disk failed, for only one redundant parity disk is needed to detect an error These extra disks are truly "redundant" since most disk controllers can already detect if a disk failed either through special signals provided in the disk interface or the extra checking information at the end of a sector used to detect and correct soft errors So information on the failed disk can be reconstructed by calculating the parity of the remaining good disks and then comparing bit-by-bit to the parity calculated for the original full

group When these two parities agree, the failed bit was a 0, otherwise it was a 1 If the check disk is the failure, just read all the data disks and store the group parity in the replacement disk

Reducing the check disks to one per group (C=1) reduces the overhead cost to between 4% and 10% for the group sizes considered here The performance for the third level RAID system is the same as the Level 2 RAID, but the effective performance per disk increases since it needs fewer check disks This reduction in total disks also increases reliability, but since it is still larger than the useful lifetime of disks, this is a minor point One advantage of a level 2 system over level 3 is that the extra check information associated with each sector to correct soft errors is not needed, increasing the capacity per disk by perhaps 10% Level 2 also allows all soft errors to be corrected "on the fly" without having to reread a sector Table IV summarizes the third level RAID characteristics and Figure 3 compares the sector layout and check disks for levels 2 and 3

MTTF		Exceeds Useful Lifetime		
		G=10 (820,000 hrs or >90 years)		*G=25* (346,000 hrs or 40 years)
Total Number of Disks		1 10D		1 04D
Overhead Cost		10%		4%
Useable Storage Capacity		91%		96%
Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk L3*	*L3/L2*	*L3/L1*
Large Reads	*D/S*	91/S	127%	91%
Large Writes	*D/S*	91/S	127%	182%
Large R-M-W	*D/S*	91/S	127%	136%
Small Reads	*D/SG*	09/S	127%	8%
Small Writes	*D/2SG*	05/S	127%	8%
Small R-M-W	*D/SG*	09/S	127%	11%

Efficiency Per Disk L3	*L3/L2*	*L3/L1*
96/S	112%	96%
96/S	112%	192%
96/S	112%	142%
04/S	112%	3%
02/S	112%	3%
04/S	112%	5%

Table IV *Characteristics of a Level 3 RAID The L3/L2 column gives the % performance of L3 in terms of L2 and the L3/L1 column gives it in terms of L1 (>100% means L3 is faster) The performance for the full systems is the same in RAID levels 2 and 3, but since there are fewer check disks the performance per disk improves*

Park and Balasubramanian proposed a third level RAID system without suggesting a particular application [Park86] Our calculations suggest it is a much better match to supercomputer applications than to transaction processing systems This year two disk manufacturers have announced level 3 RAIDs for such applications using synchronized 5 25 inch disks with G=4 and C=1 one from Maxtor and one from Micropolis [Maginnis 87]

This third level has brought the reliability overhead cost to its lowest level, so in the last two levels we improve performance of small accesses without changing cost or reliability

10. Fourth Level RAID Independent Reads/Writes

Spreading a transfer across all disks within the group has the following advantage

- Large or grouped transfer time is reduced because transfer bandwidth of the entire array can be exploited.

But it has the following disadvantages as well

- Reading/writing to a disk in a group requires reading/writing to all the disks in a group, levels 2 and 3 RAIDs can perform only one I/O at a time per group
- If the disks are not synchronized, you do not see average seek and rotational delays, the observed delays should move towards the worst case, hence the S factor in the equations above

This fourth level RAID improves performance of small transfers through parallelism--the ability to do more than one I/O per group at a time We no longer spread the individual transfer information across several disks, but keep each individual unit in a single disk

The virtue of bit-interleaving is the easy calculation of the Hamming code needed to detect or correct errors in level 2 But recall that in the third level RAID we rely on the disk controller to detect errors within a single disk sector Hence, if we store an individual transfer unit in a single sector, we can detect errors on an individual read without accessing any other disk Figure 3 shows the different ways the information is stored in a sector for RAID levels 2, 3, and 4 By storing a whole transfer unit in a sector, reads can be independent and operate at the maximum rate of a disk yet still detect errors Thus the primary change between level 3 and 4 is that we interleave data between disks at the sector level rather than at the bit level

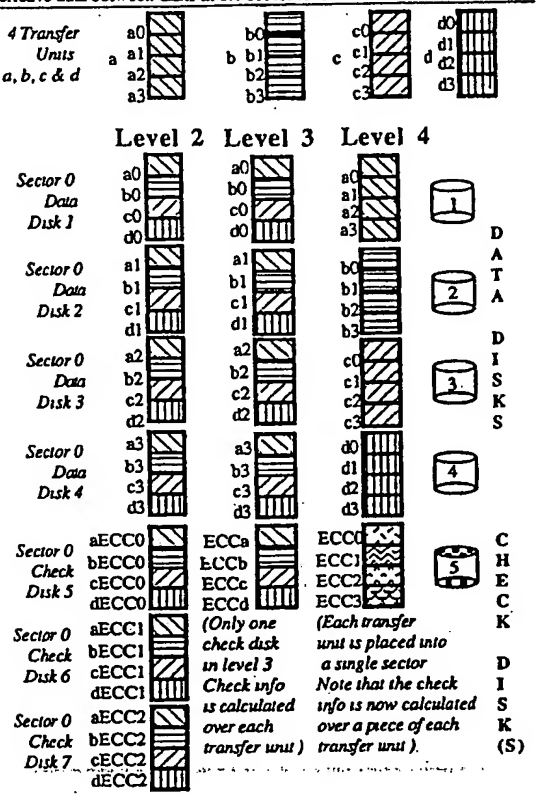


Figure 3 *Comparison of location of data and check information in sectors for RAID levels 2, 3, and 4 for G=4 Not shown is the small amount of check information per sector added by the disk controller to detect and correct soft errors within a sector Remember that we use physical sector numbers and hardware control to explain these ideas but RAID can be implemented by software using logical sectors and disks*

At first thought you might expect that an individual write to a single sector still involves all the disks in a group since (1) the check disk must be rewritten with the new parity data, and (2) the rest of the data disks must be read to be able to calculate the new parity data Recall that each parity bit is just a single exclusive OR of all the corresponding data bits in a group In level 4 RAID, unlike level 3, the parity calculation is much simpler since, if we know the old data value and the old parity value as well as the new data value, we can calculate the new parity information as follows

new parity = (old data xor new data) xor old parity

In level 4 a small write then uses 2 disks to perform 4 accesses--2 reads and 2 writes--while a small read involves only one read on one disk Table V summarizes the fourth level RAID characteristics Note that all small accesses improve--dramatically for the reads--but the small read-modify-write is still so slow relative to a level 1 RAID that its applicability to transaction processing is doubtful Recently Salem and Garcia-Molina proposed a Level 4 system [Salem 86]

Before proceeding to the next level we need to explain the performance of small writes in Table V (and hence small read-modify-writes since they entail the same operations in this RAID) The formula for the small writes divides D by 2 instead of 4 because 2

accesses can proceed in parallel the old data and old parity can be read at the same time and the new data and new parity can be written at the same time The performance of small writes is also divided by G because the single check disk in a group must be read and written with every small write in that group, thereby limiting the number of writes that can be performed at a time to the number of groups

The check disk is the bottleneck, and the final level RAID removes this bottleneck

MTTF		Exceeds Useful Lifetime					
		G=10 (820,000 hrs or >90 years)			*G=25* (346,000 hrs or 40 years)		
Total Number of Disks		1 10D			1 04D		
Overhead Cost		10%			4%		
Useable Storage Capacity		91%			96%		
Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk* *L4*	*L4/L3*	*L4/L1*	*Efficiency Per Disk* *L4*	*L4/L3*	*L4/L1*
Large Reads	*D/S*	91/S	100%	91%	96/S	100%	96%
Large Writes	*D/S*	91/S	100%	182%	96/S	100%	192%
Large R-M-W	*D/S*	91/S	100%	136%	96/S	100%	146%
Small Reads	*D*	91	1200%	91%	96	3000%	96%
Small Writes	*D/2G*	05	120%	9%	02	120%	4%
Small R-M-W	*D/G*	09	120%	14%	04	120%	6%

Table V. *Characteristics of a Level 4 RAID The L4/L3 column gives the % performance of L4 in terms of L3 and the L4/L1 column gives it in terms of L1 (>100% means L4 is faster) Small reads improve because they no longer tie up a whole group at a time Small writes and R-M-Ws improve some because we make the same assumptions as we made in Table II the slowdown for two related I/Os can be ignored because only two disks are involved*

11. Fifth Level RAID: No Single Check Disk

While level 4 RAID achieved parallelism for reads, writes are still limited to one per group since every write must read and write the check disk The final level RAID distributes the data and check information across all the disks--including the check disks Figure 4 compares the location of check information in the sectors of disks for levels 4 and 5 RAIDs

The performance impact of this small change is large since RAID level 5 can support multiple individual writes per group For example, suppose in Figure 4 above we want to write sector 0 of disk 2 and sector 1 of disk 3 As shown on the left Figure 4, in RAID level 4 these writes must be sequential since both sector 0 and sector 1 of disk 5 must be written However, as shown on the right, in RAID level 5 the writes can proceed in parallel since a write to sector 0 of disk 2 still involves a write to disk 5 but a write to sector 1 of disk 3 involves a write to disk 4

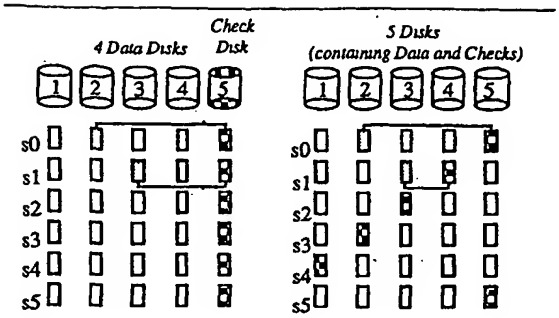
These changes bring RAID level 5 near the best of both worlds small read-modify-writes now perform close to the speed per disk of a level 1 RAID while keeping the large transfer performance per disk and high useful storage capacity percentage of the RAID levels 3 and 4 Spreading the data across all disks even improves the performance of small reads, since there is one more disk per group that contains data Table VI summarizes the characteristics of this RAID

Keeping in mind the caveats given earlier, a Level 5 RAID appears very attractive if you want to do just supercomputer applications, or just transaction processing when storage capacity is limited, or if you want to do both supercomputer applications *and* transaction processing

12. Discussion

Before concluding the paper, we wish to note a few more interesting points about RAIDs The first is that while the schemes for disk striping and parity support were presented as if they were done by hardware, there is no necessity to do so We just give the method, and the decision between hardware and software solutions is strictly one of cost and benefit. For example, in cases where disk buffering is effective, there is no extra disks reads for level 5 small writes since the old data and old parity would be in main memory, so software would give the best performance as well as the least cost.

In this paper we have assumed the transfer unit is a multiple of the sector As the size of the smallest transfer unit grows larger than one



(a) Check information for Level 4 RAID for G=4 and C=1 The sectors are shown below the disks (The checked areas indicate the check information) Writes to s0 of disk 2 and s1 of disk 3 imply writes to s0 and s1 of disk 5 The check disk (5) becomes the write bottleneck.

(b) Check information for Level 5 RAID for G=4 and C=1 The sectors are shown below the disks, with the check information and data spread evenly through all the disks Writes to s0 of disk 2 and s1 of disk 3 still imply 2 writes, but they can be split across 2 disks to s0 of disk 5 and to s1 of disk 4

Figure 4 *Location of check information per sector for Level 4 RAID vs. Level 5 RAID*

MTTF		Exceeds Useful Lifetime					
		G=10 (820,000 hrs or >90 years)			*G=25* (346,000 hrs or 40 years)		
Total Number of Disks		1 10D			1 04D		
Overhead Cost		10%			4%		
Useable Storage Capacity		91%			96%		
Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk* *L5*	*L5/L4*	*L5/L1*	*Efficiency Per Disk* *L5*	*L5/L4*	*L5/L1*
Large Reads	*D/S*	91/S	100%	91%	96/S	100%	96%
Large Writes	*D/S*	91/S	100%	182%	96/S	100%	192%
Large R-M-W	*D/S*	91/S	100%	136%	96/S	100%	144%
Small Reads	*(1+C/G)D*	1 00	110%	100%	1 00	104%	100%
Small Writes	*(1+C/G)D/4*	25	550%	50%	25	1300%	50%
Small R-M-W	*(1+C/G)D/2*	50	550%	75%	50	1300%	75%

Table VI *Characteristics of a Level 5 RAID The L5/L4 column gives the % performance of L5 in terms of L4 and the L5/L1 column gives it in terms of L1 (>100% means L5 is faster) Because reads can be spread over all disks, including what were check disks in level 4, all small I/Os improve by a factor of 1+C/G Small writes and R-M-Ws improve because they are no longer constrained by group size, getting the full disk bandwidth for the 4 I/O's associated with these accesses We again make the same assumptions as we made in Tables II and V the slowdown for two related I/Os can be ignored because only two disks are involved*

sector per drive--such as a full track with an I/O protocol that supports data returned out-of-order--then the performance of RAIDs improves significantly because of the full track buffer in every disk For example, if every disk begins transferring to its buffer as soon as it reaches the next sector, then S may reduce to less than 1 since there would be virtually no rotational delay With transfer units the size of a track, it is not even clear if synchronizing the disks in a group improves RAID performance

This paper makes two separable points the advantages of building I/O systems from personal computer disks and the advantages of five different disk array organizations, independent of disks used in those array The later point starts with the traditional mirrored disks to achieve acceptable reliability, with each succeeding level improving

- *the data rate*, characterized by a small number of requests per second for massive amounts of sequential information (supercomputer applications),

- *the I/O rate*, characterized by a large number of read-modify-writes to a small amount of random information (transaction-processing),
- or *the useable storage capacity*,

or possibly all three

Figure 5 shows the performance improvements per disk for each level RAID The highest performance per disk comes from either Level 1 or Level 5 In transaction-processing situations using no more than 50% of storage capacity, then the choice is mirrored disks (Level 1) However, if the situation calls for using more than 50% of storage capacity, or for supercomputer applications, or for combined supercomputer applications and transaction processing, then Level 5 looks best Both the strength and weakness of Level 1 is that it duplicates data rather than calculating check information, for the duplicated data improves read performance but lowers capacity and write performance,while check data is useful only on a failure

Inspired by the space-time product of paging studies [Denning 78], we propose a single figure of merit called the *space-speed product* the useable storage fraction times the efficiency per event Using this metric, Level 5 has an advantage over Level 1 of 1 7 for reads and 3 3 for writes for G=10

Let us return to the first point, the advantages of building I/O system from personal computer disks Compared to traditional Single Large Expensive Disks (SLED), Redundant Arrays of Inexpensive Disks (RAID) offer significant advantages for the same cost Table VII compares a level 5 RAID using 100 inexpensive data disks with a group size of 10 to the IBM 3380 As you can see, a level 5 RAID offers a factor of roughly 10 improvement in performance, reliability, and power consumption (and hence air conditioning costs) and a factor of 3 reduction in size over this SLED Table VII also compares a level 5 RAID using 10 inexpensive data disks with a group size of 10 to a Fujitsu M2361A "Super Eagle" In this comparison RAID offers roughly a factor of 5 improvement in performance, power consumption, and size with more than two orders of magnitude improvement in (calculated) reliability

RAID offers the further advantage of modular growth over SLED Rather than being limited to 7,500 MB per increase for $100,000 as in the case of this model of IBM disk, RAIDs can grow at either the group size (1000 MB for $11,000) or, if partial groups are allowed, at the disk size (100 MB for $1,100) The flip side of the coin is that RAID also makes sense in systems considerably smaller than a SLED Small incremental costs also makes hot standby spares practical to further reduce MTTR and thereby increase the MTTF of a large system For example, a 1000 disk level 5 RAID with a group size of 10 and a few standby spares could have a calculated MTTF of over 45 years

A final comment concerns the prospect of designing a complete transaction processing system from either a Level 1 or Level 5 RAID The drastically lower power per megabyte of inexpensive disks allows systems designers to consider battery backup for the whole disk array--the power needed for 110 PC disks is less than two Fujitsu Super Eagles Another approach would be to use a few such disks to save the contents of battery backed-up main memory in the event of an extended power failure The smaller capacity of these disks also ties up less of the database during reconstruction, leading to higher availability (Note that Level 5 ties up all the disks in a group in event of failure while Level 1 only needs the single mirrored disk during reconstruction, giving Level 1 the edge in availability)

13. Conclusion

RAIDs offer a cost effective option to meet the challenge of exponential growth in the processor and memory speeds We believe the size reduction of personal computer disks is a key to the success of disk arrays, just as Gordon Bell argues that the size reduction of microprocessors is a key to the success in multiprocessors [Bell 85] In both cases the smaller size simplifies the interconnection of the many components as well as packaging and cabling While large arrays of mainframe processors (or SLEDs) are possible, it is certainly easier to construct an array from the same number of microprocessors (or PC drives) Just as Bell coined the term "multi" to distinguish a multiprocessor made from microprocessors, we use the term "RAID" to identify a disk array made from personal computer disks

With advantages in cost-performance, reliability, power consumption, and modular growth, we expect RAIDs to replace SLEDs in future I/O systems There are, however, several open issues that may bare on the practicality of RAIDs

- *What is the impact of a RAID on latency?*
- *What is the impact on MTTF calculations of non-exponential failure assumptions for individual disks?*
- *What will be the real lifetime of a RAID vs calculated MTTF using the independent failure model?*
- *How would synchronized disks affect level 4 and 5 RAID performance?*
- *How does "slowdown" S actually behave? [Livny 87]*
- *How do defective sectors affect RAID?*
- *How do you schedule I/O to level 5 RAIDs to maximize write parallelism?*
- *Is there locality of reference of disk accesses in transaction processing?*
- *Can information be automatically redistributed over 100 to 1000 disks to reduce contention?*
- *Will disk controller design limit RAID performance?*
- *How should 100 to 1000 disks be constructed and physically connected to the processor?*
- *What is the impact of cabling on cost, performance, and reliability?*
- *Where should a RAID be connected to a CPU so as not to limit performance? Memory bus? I/O bus? Cache?*
- *Can a file system allow differ striping policies for different files?*
- *What is the role of solid state disks and WORMs in a RAID?*
- *What is the impact on RAID of "parallel access" disks (access to every surface under the read/write head in parallel)?*

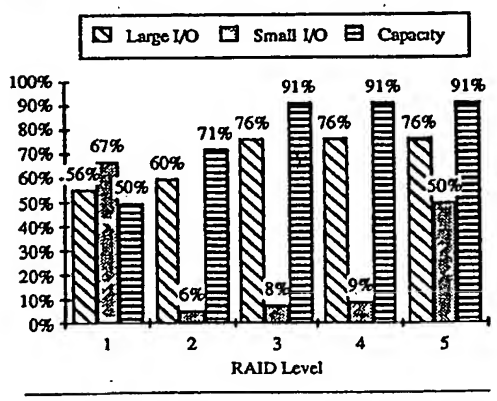


Figure 5 *Plot of Large (Grouped) and Small (Individual) Read-Modify-Writes per second per disk and useable storage capacity for all five levels of RAID (D=100, G=10) We assume a single S factor uniformly for all levels with S=1 3 where it is needed*

Characteristics	*RAID 5L (100 10) (CP3100)*	*SLED (IBM 3380)*	*RAID v SLED (>1 better for RAID)*	*RAID 5L (10,10) (CP3100)*	*SLED (Fujitsu M2361)*	*RAID v SLED (>1 better for RAID)*
Formatted Data Capacity (MB)	10,000	7,500	1 33	1,000	600	1 67
Price/MB (controller incl)	$11-$8	$18-$10	2 2- 9	$11-$8	$20-$17	2 5-1 5
Rated MTTF (hours)	820,000	30,000	27 3	8,200,000	20,000	410
MTTF in practice (hours)	?	100,000	?	?	?	?
No Actuators	110	4	22 5	11	1	11
Max I/O's/Actuator	30	50	6	30	40	8
Max Grouped RMW/box	1250	100	12 5	125	20	6 2
Max Individual RMW/box	825	100	8 2	83	20	4 2
Typ I/O's/Actuator	20	30	7	20	24	8
Typ Grouped RMW/box	833	60	13 9	83	12	6 9
Typ Individual RMW/box	550	60	9 2	55	12	4 6
Volume/Box (cubic feet)	10	24	2 4	1	3 4	3 4
Power/box (W)	1100	6,600	6 0	110	640	5 8
Min Expansion Size (MB)	100-1000	7,500	7 5-75	100-1000	600	0 6-6

Table VII *Comparison of IBM 3380 disk model AK4 to Level 5 RAID using 100 Conners & Associates CP 3100s disks and a group size of 10 and a comparison of the Fujitsu M2361A "Super Eagle" to a level 5 RAID using 10 inexpensive data disks with a group size of 10 Numbers greater than 1 in the comparison columns favor the RAID*

Acknowledgements

We wish to acknowledge the following people who participated in the discussions from which these ideas emerged Michael Stonebraker, John Ousterhout, Doug Johnson, Ken Lutz, Anapum Bhide, Gaetano Bonello Mark Hill, David Wood, and students in SPATS seminar offered at U C Berkeley in Fall 1987 We also wish to thank the following people who gave comments useful in the preparation of this paper Anapum Bhide, Pete Chen, Ron David, Dave Ditzel, Fred Douglis, Dieter Gawlick, Jim Gray, Mark Hill Doug Johnson, Joan Pendleton, Martin Schulze, and Hervé Touati This work was supported by the National Science Foundation under grant # MIP-8715235

Appendix Reliability Calculation

Using probability theory we can calculate the $MTTF_{Group}$ We first assume independent and exponential failure rates Our model uses a biased coin with the probability of heads being the probability that a second failure will occur within the MTTR of a first failure Since disk failures are exponential

$$\text{Probability(at least one of the remaining disks failing in MTTR)} = [\,1 - (e^{-MTTR/MTTF_{Disk}})^{(G+C-1)}\,]$$

In all practical cases

$$MTTR << \frac{MTTF_{Disk}}{G+C}$$

and since $(1 - e^{-X})$ is approximately X for $0 < X << 1$

$$\text{Probability(at least one of the remaining disks failing in MTTR)} = MTTR*(G+C-1)/MTTF_{Disk}$$

Then that on a disk failure we flip this coin

heads => a system crash, because a second failure occurs before the first was repaired,

tails => recover from error and continue

Then

$$MTTF_{Group} = \text{Expected[Time between Failures]} * \text{Expected[no of flips until first heads]}$$

$$= \frac{\text{Expected[Time between Failures]}}{\text{Probability(heads)}}$$

$$= \frac{MTTF_{Disk}}{(G+C)*(MTTR*(G+C-1)/MTTF_{Disk})}$$

$$MTTF_{Group} = \frac{(MTTF_{Disk})^2}{(G+C)*(G+C-1)*MTTR}$$

Group failure is not precisely exponential in our model, but we have validated this simplifying assumption for practical cases of MTTR << MTTF/(G+C) This makes the MTTF of the whole system just $MTTF_{Group}$ divided by the number of groups, n_G

References

[Bell 84] C G Bell, "The Mini and Micro Industries," *IEEE Computer* Vol 17 No 10 (October 1984), pp 14-30

[Joy 85] B Joy presentation at ISSCC '85 panel session, Feb 1985

[Siewiorek 82] D P Siewiorek, C G Bell, and A Newell, *Computer Structures: Principles and Examples*, p 46

[Moore 75] G E Moore, "Progress in Digital Integrated Electronics," *Proc IEEE Digital Integrated Electronic Device Meeting*, (1975), p 11

[Myers 86] G J Myers A Y C Yu, and D L House, "Microprocessor Technology Trends " *Proc IEEE*, Vol 74, no 12, (December 1986), pp 1605-1622

[Garcia 84] H Garcia Molina, R Cullingford, P Honeyman, R Lipton, "The Case for Massive Memory," Technical Report 326, Dept of EE and CS, Princeton Univ , May 1984

[Myers 86] W Myers, "The Competitiveness of the United States Disk Industry," *IEEE Computer*, Vol 19, No 11 (January 1986), pp 85-90

[Frank 87] P D Frank, "Advances in Head Technology," presentation at *Challenges in Disk Technology Short Course*, Institute for Information Storage Technology, Santa Clara University, Santa Clara, California, December 15-17, 1987

[Stevens 81] L D Stevens, "The Evolution of Magnetic Storage," *IBM Journal of Research and Development*, Vol 25, No 5, Sept 1981, pp 663-675

[Harker 81] J M Harker *et al* , "A Quarter Century of Disk File Innovation," *ibid* , pp 677-689

[Amdahl 67] G M Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," *Proceedings AFIPS 1967 Spring Joint Computer Conference* Vol 30 (Atlantic City, New Jersey April 1967), pp 483-485

[Boral 83] H Boral and D J DeWitt, "Database Machines An Ideas Whose Time Has Passed? A Critique of the Future of Database Machines," *Proc International Conf on Database Machines*, Edited by H -O Leilich and M Misskoff, Springer-Verlag, Berlin, 1983

[IBM 87] "IBM 3380 Direct Access Storage Introduction," IBM GC 26-4491-0, September 1987

[Gawlick 87] D Gawlick, private communication, Nov , 1987

[Fujitsu 87] "M2361A Mini-Disk Drive Engineering Specifications," (revised) Feb , 1987, B03P-4825-0001A

[Adaptec 87] AIC-6250, *IC Product Guide*, Adaptec, stock # DB0003-00 rev B, 1987, p 46

[Livny 87] Livny, M , S Khoshafian, H Boral, "Multi-disk management algorithms," *Proc of ACM SIGMETRICS*, May 1987

[Kim 86] M Y Kim, "Synchronized disk interleaving," *IEEE Trans on Computers*, vol C-35, no 11, Nov 1986

[Salem 86] K Salem and Garcia-Molina, H , "Disk Striping," *IEEE 1986 Int Conf on Data Engineering*, 1986

[Bitton 88] D Bitton and J Gray, "Disk Shadowing," *in press*, 1988

[Kurzweil 88] F Kurzweil, "Small Disk Arrays - The Emerging Approach to High Performance," presentation at Spring COMPCON 88, March 1, 1988, San Francisco, CA

[Hamming 50] R W Hamming, "Error Detecting and Correcting Codes," *The Bell System Technical Journal*, Vol XXVI, No 2 (April 1950), pp 147-160

[Hillis 87] D Hillis, private communication, October, 1987

[Park 86] A Park and K Balasubramanian, "Providing Fault Tolerance in Parallel Secondary Storage Systems," Department of Computer Science, Princeton University, CS-TR-057-86, Nov 7, 1986

[Maginnis 87] N B Maginnis, "Store More, Spend Less Mid-range Options Abound,"*Computerworld*, Nov 16, 1987, p 71

[Denning 78] P J Denning and D F Slutz, "Generalized Working Sets for Segment Reference Strings," *CACM*, vol 21, no 9, (Sept 1978) pp 750-759

[Bell 85] Bell, C G , "Multis a new class of multiprocessor computers,"*Science*, vol 228 (April 26, 1985) 462-467

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-293314
(43)Date of publication of application : 20.10.2000

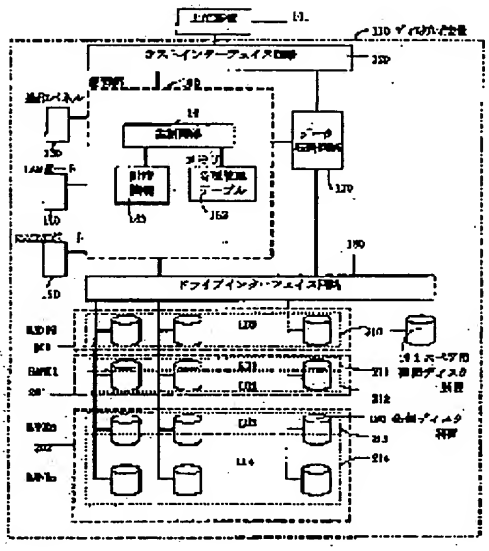
(51)Int.Cl. G06F 3/06
G06F 1/32
G11B 19/02

(21)Application number : 11-097290
(22)Date of filing : 05.04.1999

(71)Applicant : HITACHI LTD
(72)Inventor : HAKAMATA KAZUO
TAKAMOTO KENICHI
KOBAYASHI MASAAKI

(54) DISK ARRAY DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To suppress the power consumption of a magnetic disk drive mounted on a disk array device.
SOLUTION: This device is provided with a means which controls the relation between the configuration of plural magnetic disk drives and access from a host device 101, a power-saving controlling means which controls the power- saving (selection of power on/off and power-saving mode) of magnetic disk drives in a set logical drive and a controlling means which controls the diagnoses of the magnetic disk drives. This disk array device 110 shifts a prescribed magnetic disk drive to a power-saving mode or turns off the power (power- saving processing) after access from the device 101 does not exist any more and a predetermined time elapses. The magnetic disk drive undergoing power- save processing is subjected to diagnosis execution after a prescribed time passes at the start of the power-saving processing or when a designated time comes in order to maintain its reliability.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－293314
（P2000－293314A）
(43)公開日　平成12年10月20日(2000.10.20)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)	
Ｇ０６Ｆ	3/06	３０１	Ｇ０６Ｆ	3/06	３０１Ａ	５Ｂ０１１
		５４０			５４０	５Ｂ０６５
	1/32		Ｇ１１Ｂ	19/02	５０１Ｆ	５Ｄ０６６
Ｇ１１Ｂ	19/02	５０１	Ｇ０６Ｆ	1/00	３３２Ｚ	

審査請求　未請求　請求項の数15　ＯＬ　（全 12 頁）

(21)出願番号　特願平11－97290

(22)出願日　平成11年４月５日(1999.4.5)

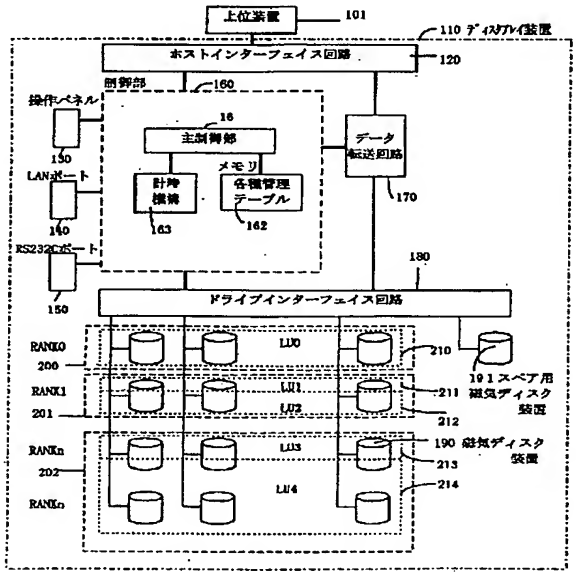
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　袴田　和夫
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(72)発明者　▲高▼本　實一
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　ディスクアレイ装置

(57)【要約】　（修正有）
【課題】ディスクアレイ装置に実装されている、磁気ディスク装置の消費電力を抑える。
【解決手段】複数の磁気ディスク装置の構成と上位装置からのアクセスとの関連を制御する手段と、設定された論理ドライブ内の磁気ディスク装置の節電（電源オンオフや節電モードの選択）を制御する節電制御手段と、磁気ディスク装置の診断を制御する制御手段を設ける。ディスクアレイ装置において、所定の磁気ディスク装置に対し、上位装置からアクセスが無くなり予め定めた時間経過後、節電モードに移行させるか又は、電源をオフにする（節電処理）。節電処理をした磁気ディスク装置は、信頼性を維持する為に、節電処理の開始から所定時間経過後、または、指定の時刻になった時、診断を実行する。

図１



【特許請求の範囲】

【請求項１】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

２以上の磁気ディスク装置であって、電子回路の一部の電力消費を抑制する第１の節電モードと、スピンドルの回転を制御することにより電力消費を抑制する第２の節電モードとを有する磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記磁気ディスク装置の組に対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を第２の節電モードとすることを特徴とするディスクアレイ装置。

【請求項２】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

２以上の磁気ディスク装置であって、電子回路の一部の電力消費を抑制する第１の節電モードと、スピンドルの回転を制御することにより電力消費を抑制する第２の節電モードとを有する磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記磁気ディスク装置の組に対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を第１の節電モードとすることを特徴とするディスクアレイ装置。

【請求項３】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

節電モードを有する２以上の磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組の１つに２以上の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記２以上の論理ユニットの全てに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとすることを特徴とするディスクアレイ装置。

【請求項４】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

節電モードを有する２以上の磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組を超えて１の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記１の論理ユニットに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとすることを特徴とするディスクアレイ装置。

【請求項５】請求項３又は請求項４記載のディスクアレイ装置において、前記節電モードは、スピンドルの回転を制御することにより電力消費を抑制するモードであるディスクアレイ装置。

【請求項６】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

節電モードを有する２以上の磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能と、磁気ディスク装置の診断を行う機能とを有する制御部とを有し、

前記磁気ディスク装置の組に対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとし、所定時間経過後又は指定した時刻に、該節電モードにあった磁気ディスク装置の組に対し、診断を行うことを特徴とするディスクアレイ装置。

【請求項７】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

節電モードを有する２以上の磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組の１つに２以上の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記２以上の論理ユニットの全てに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとし、所定時間経過後又は指定した時刻に、該節電モードにあった磁気ディスク装置の組に対し、診断を行うことを特徴とするディスクアレイ装置。

【請求項８】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、

節電モードを有する２以上の磁気ディスク装置と、

前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組を超えて１の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能とを有する制御部とを有し、

前記１の論理ユニットに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとし、所定時間経過後又は指定した時刻に、該節電モードにあった磁気ディスク装置の組に対し、診断を行うことを特徴とするディスクアレイ装置。

【請求項９】請求項１乃至請求項８記載のいずれか１の

ディスクアレイ装置において、前記所定のアクセスが無い場合の所定の時間を、上位装置から指定するディスクアレイ装置。

【請求項１０】請求項６乃至請求項８記載のいずれか１のディスクアレイ装置において、前記所定時間経過後の所定時間又は前記時刻の指定を、上位装置から指定するディスクアレイ装置。

【請求項１１】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、
節電モードを有する２以上の磁気ディスク装置と、
前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組の１つに２以上の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能と、ＬＡＮ若しくはＲＳ２３２Ｃを制御する機能とを有する制御部とを有し、
前記２以上の論理ユニットの全てに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとすることの指定を、前記ＬＡＮ若しくはＲＳ２３２Ｃを経由して行うことを特徴とするディスクアレイ装置。

【請求項１２】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、
節電モードを有する２以上の磁気ディスク装置と、
前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組を超えて１の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能と、ＬＡＮ若しくはＲＳ２３２Ｃを制御する機能とを有する制御部とを有し、
前記１の論理ユニットに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとすることの指定を、前記ＬＡＮ若しくはＲＳ２３２Ｃを経由して行うことを特徴とするディスクアレイ装置。

【請求項１３】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、
節電モードを有する２以上の磁気ディスク装置と、
前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組の１つに２以上の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能と、ＬＡＮ若しくはＲＳ２３２Ｃを制御する機能とを有する制御部とを有し、
前記２以上の論理ユニットの全てに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとし、所定時間経過後又は指定した時刻に、該節電モードにあった磁気ディスク装置の組に対し、診断を行うことを特徴とするディスクアレイ装置。

【請求項１４】上位装置に接続され、該上位装置との間で情報の授受を行うディスクアレイ装置において、
節電モードを有する２以上の磁気ディスク装置と、
前記磁気ディスク装置の組と上位装置からのアクセスとを対応付ける機能と、該磁気ディスク装置の組を超えて１の論理ユニットを対応させる機能と、上位装置からのアクセスから次のアクセスまでの時間を計数する機能と、１の磁気ディスク装置を特定してその節電モードを実行させる機能と、ＬＡＮ若しくはＲＳ２３２Ｃを制御する機能とを有する制御部とを有し、
前記１の論理ユニットに対し、所定の時間アクセスが無い場合に、前記磁気ディスク装置を節電モードとし、所定時間経過後又は指定した時刻に、該節電モードにあった磁気ディスク装置の組に対し、診断を行うことを特徴とするディスクアレイ装置。

【請求項１５】請求項１３及び請求項１４記載のいずれか１のディスクアレイ装置において、前記所定時間経過後の所定時間又は前記時刻の指定を、前記ＬＡＮ若しくはＲＳ２３２Ｃを経由して行うことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、コンピュータ装置、周辺機器その他の電子装置の節電に関し、特に、コンピュータ装置に接続されるディスクアレイ装置の省エネルギー化に関する。

【０００２】

【従来の技術】電子装置には高機能化、低価格化の要求が定常的にある。最近、環境問題も重視されるようになり、省エネルギー化も要求されるようになってきた。

【０００３】ディスクアレイ装置は、コンピュータ装置の周辺機器であり、コンピュータのデータを格納する外部記憶装置として用いられている。ディスクアレイ装置には、複数の磁気ディスク装置が搭載され、ディスクアレイ装置及びこの上位装置であるコンピュータシステムが機能しているときは、通常は、ディスクアレイ装置に搭載された複数の磁気ディスク装置の全てが動作中となっている。

【０００４】これら磁気ディスク装置は、上位装置からのアクセスが発生したとき、制御情報やデータの送信又は受信が行えるように、アクセス対象の磁気ディスク装置が動作中でなければならない。一方、所定の時間、上位装置や制御装置からアクセスがない場合には、磁気ディスク装置を「動作中」にしておく必要はない。

【０００５】ここで、「動作中」とは、制御装置や上位装置から、その磁気ディスク装置に対して制御情報やデータが入力された場合に、その磁気ディスク装置が即座に応答する状態にあることをいう。いわゆる磁気ディスク装置がスリープモードにあって、磁気ディスク媒体が

静止している状態から通常の回転数までスピンアップされる時間の経過後に応答することを含まない。また、磁気ディスク装置が節電モードにあって、その節電モードを解除した後に応答することを含まない。

【０００６】また、アクセスとは、磁気ディスク装置から見て自己を対象とするコマンドが発行されたか否か、自己を対象とする情報の授受が要求されたか否かを意味する。通常は、上位装置の情報取得又は格納の命令に対応して、ドライブインタフェースを含むこれらより上位装置側の電子回路から、対象となる磁気ディスク装置へ信号が発行されることに対応している。

【０００７】磁気ディスク装置には、一般に、上位装置からのアクセスを受領したときに電源をオンとし、所定の電子回路をオンとして、磁気記録媒体の回転を行い、アクセスに応答（情報の送信又は受信）を行い、その終了後に、所定の電子回路や磁気記録媒体の回転のためのモータの電源を、段階的にオフとする制御方法がある。つまり、磁気ディスク装置は、種々の節電モードを内蔵しており、主として上位装置又は制御装置からのアクセス頻度に応じて、所定の節電モードを選択して、そのモードに自動的に移行する機能を有している。

【０００８】従来のディスクアレイ装置では、実装されている磁気ディスク装置又は上位装置から認識可能に設定されている磁気ディスク装置は、ディスクアレイ装置の起動によって動作中となるよう制御されていた。ディスクアレイ装置が起動した後は、この電源をオフとするまで、これに搭載されている上記の磁気ディスク装置は動作中となっている。ディスクアレイ装置の電源をオフとする操作がなされると、これを契機に磁気ディスク装置を「動作中」から電源遮断とするシーケンスが働き、その後、ディスクアレイ装置の電源が遮断される。

【０００９】上位装置から認識可能に設定されていない磁気ディスク装置、及び、スペア用に実装されている磁気ディスク装置は、上位装置から使用可能な設定をしたとき、又は、スペア用の磁気ディスク装置が使用されるときに、それぞれ、電源を投入し動作中とする。一度、動作中とした磁気ディスク装置は、ディスクアレイ装置の電源を遮断するまでは、個別に電源を遮断することはなかった。このため磁気ディスク装置の台数が増加すると、例えば、数百台の磁気ディスク装置を接続するディスクアレイ装置となると、節電対策が必須となる。

【００１０】

【発明が解決しようとする課題】ディスクアレイ装置では、上位装置からのアクセスが少ない場合には、全ての磁気ディスク装置を動作中にしておく必要がない。このため上記の制御方法を適用すれば省電力化が可能であると考えられる。しかしアクセス受領を契機に磁気ディスク装置を「動作中」とするための時間が必要となり、ディスクアレイ装置全体としての性能は著しく低下する。性能の低下を押さえる為には、「動作中」となるまでの復帰時間の短い節電モードを選択して実行するか、スピンアップまでの時間の短い磁気ディスク装置を用いるか、磁気ディスク装置の代わりに、そのデータを保持するメモリへのアクセスを考慮しなければならない。

【００１１】また、１組の磁気ディスク装置（物理ユニット）を１つの論理ユニットとして扱わないディスクアレイ装置においては、単に、磁気ディスク装置の既存の節電モードを転用しただけでは、ディスクアレイ装置の節電を実現できない。

【００１２】

【課題を解決するための手段】１組の磁気ディスク装置群（物理ユニットグループ）毎に、上位装置又は制御装置から所定時間のアクセスが無い場合に、複数の節電モードの１つを選択して、その１組の磁気ディスク装置群を選択した節電モードとする。この節電モードには磁気記録媒体の回転を静止させるモードが含まれる。

【００１３】１組の磁気ディスク装置群（物理ユニットグループ）が２以上のロジカルユニットに対応している場合には、その２以上のロジカルユニットすべてに対し、上位装置又は制御装置から所定時間のアクセスが無い場合に、複数の節電モードの１つを選択して、その１組の磁気ディスク装置群を選択した節電モードとする。

【００１４】１組の磁気ディスク装置群（物理ユニットグループ）を超えて、１つのロジカルユニットが定義されている場合には、その１つのロジカルユニットに対応するすべての磁気ディスク装置に対し、上位装置又は制御装置から所定時間のアクセスが無い場合に、複数の節電モードの１つを選択して、その１組の磁気ディスク装置群を選択した節電モードとする。

【００１５】このような制御を実行するため、磁気ディスク装置の構成と上位装置からのアクセスとを対応付ける手段と、ディスクアレイ装置に認識される磁気ディスク装置の節電モードを選択する節電制御手段と、磁気ディスク装置の状態を診断する診断手段を設ける。ここで診断とは、磁気ディスク装置が使用可能な状態に有るか否か、その動作の確認をすることをいう。例えば、オンラインヴェリファイコマンドを実行して、そのコマンドが正常終了するか否かを確認する。

【００１６】

【発明の実施の形態】以下、本発明のディスクアレイ装置の実施例について説明する。

【００１７】図１に、上位装置１０１に接続されたディスクアレイ装置１１０の内部構成の一例を示す。

【００１８】上位装置１０１は、情報の読み書きを制御し、ディスクアレイ装置１１０に対してコマンドを発行することにより、情報の読み書きを実現する。

【００１９】ディスクアレイ装置１１０は、ホストインタフェース回路１２０、制御部１６０、データ転送回路１７０、ドライブインタフェース回路１８０、磁気ディスク装置１９０、スペア用磁気ディスク装置１９１、操

作パネル１３０、ＬＡＮ制御部１４０、および、ＲＳ２３２Ｃ制御部１５０より構成される。

【００２０】制御部１６０は、マイクロプロセッサと制御用ファームウェアにより実現される。装置全体を制御する主制御部１６１、時間を管理する計時機構１６３からなる。制御部１６０上のメモリ１６２には各種管理テーブルを置く。

【００２１】磁気ディスク装置１９０およびスペア用磁気ディスク装置１９１には、汎用部品である小型の磁気ディスク装置を用いると製造原価低減の効果が大きい。磁気ディスク装置１９０は、ＲＡＩＤ（Redundant Array Inexpensive Disks）構成となるよう、アレイ状に配置する（図１）。

【００２２】ＲＡＩＤグループは、１列（２００、２０１）または複数列（２０２）にて構成される。ＲＡＩＤグループを構成する磁気ディスク装置１９０は、上位装置１０１からアクセス可能とする為に、ＲＡＮＫなる概念を導入し、ＲＡＩＤグループ構成毎にロジカルユニットを設定する。各々の同一ＲＡＮＫ内で磁気ディスク装置の格納エリアを分割する（領域分割）。

【００２３】一般に、ディスクアレイでは複数の磁気ディスク装置に対し、例えばユーザデータを適当な大きさのデータに分割して（ストライピング）、各磁気ディスク装置に振り分けて格納する。そして、ディスクアレイが自己の磁気ディスク装置に対し、データの格納やアクセスを均等に行うことが好ましい。このためにＲＡＮＫなる概念を、論理的な複数の磁気ディスク装置から成る構成に対応させた。従って、一組の磁気ディスク装置が複数の又は１に満たない論理的ディスク装置に対応できる。

【００２４】ＲＡＮＫとロジカルユニットとの対応付けは、ディスクアレイ装置１１０の構成情報の設定により定める。ｎＲＡＮＫ＝１ロジカルユニット、１ＲＡＮＫ＝ｎロジカルユニット、または、ｍＲＡＮＫ＝ｎロジカルユニットとすることが可能である。図１では、ＲＡＮＫ０（２００）にＬＵ０（２１０、ロジカルユニット番号０）、ＲＡＮＫ１（２０１）にＬＵ１（２１１）とＬＵ２（２１２）、ＲＡＮＫｎ（２０２）にＬＵ３（２１３）とＬＵ４（２１４）の設定である。

【００２５】スペア用磁気ディスク装置１９１は、ＲＡＩＤグループを構成した磁気ディスク装置１９０に障害が発生したとき、代替用の磁気ディスク装置として使用する。具体的には、診断を実行して障害を検出し、障害のある磁気ディスク装置１９１とスペア用磁気ディスク装置１９１を置換する。図１ではスペア用磁気ディスク装置は１台のみが図示されているが、これに限られない。

【００２６】図２は、メモリ１６２上の各種管理テーブル構造を示す。

【００２７】磁気ディスク装置管理テーブル２５０は、設定したロジカルユニット毎の情報を管理する。設定したロジカルユニットは、その番号を設定ＬＵＮに登録することにより管理される。ＲＡＮＫｎ（２０２）の構成の場合は、複数列構成であることから、管理ＬＵＮを用いて、ここに分割の枝番号を登録して管理する。また、上位装置１０１からロジカルユニットに対してアクセスを受領した時刻をアクセス時刻に登録する。設定したロジカルユニット内の磁気ディスク装置１９０の位置を、磁気ディスク装置位置に登録する。ＲＡＮＫｎ（２０２）に示すように同一のＲＡＮＫ内に複数のロジカルユニットを設定した場合は、設定した複数の他のロジカルユニット番号を関連ＬＵＮに登録する。磁気ディスク装置１９０を節電モードとした時刻又は磁気ディスク装置１９０の電源を遮断した時刻を、節電開始時刻に登録する。

【００２８】節電待ち時間２６０は、上位装置１０１からの最後のアクセスを受領した後、磁気ディスク装置１９０が節電モードとなるまで又はその電源を遮断するまでの時間を登録する。

【００２９】診断開始時間２７０は、磁気ディスク装置１９０が節電モード等となってから磁気ディスク装置１９０の診断を開始するまでの時間、または、節電モード等となった磁気ディスク装置１９０の診断を実行する時間を登録する。

【００３０】以下、本発明に関する節電方法をフローチャート（図３、図４および図５）を用いて説明する。

【００３１】上位装置１０１は、アプリケーションの実行を行い、必要な情報はディスクアレイ装置１１０に格納する。必要な情報は、ディスクアレイ装置に対してリード又はライトのアクセスにより取り出し又は書き込み、必要な機能をアクティブにする。

【００３２】ディスクアレイ装置１１０は、その起動の際に、上位装置からのアクセスに応答可能とするため、必要な磁気ディスク装置１９０の電源を投入する。

【００３３】スペア用磁気ディスク装置１９１は、通常は上位装置１０１からアクセスされることがないため、その電源が投入され正常に動作することが確認された後、電源が遮断される。尚、スペア用磁気ディスク装置１９１の電源を遮断する代わりに、節電モードで待機させても良い。

【００３４】ディスクアレイ装置１１０は、磁気ディスク装置１９０の電源を投入し、磁気ディスク装置１９０の正常動作を確認した後、記憶装置管理テーブル２５０の、設定ＬＵＮ、管理ＬＵＮ、アクセス時刻（磁気ディスク装置１９０の正常を確認した時刻）、磁気ディスク装置位置、および、関連ＬＵＮを登録する（図３）。管理ＬＵＮは、ＲＡＮＫｎ（２０２）が、複数の磁気ディスク装置１９０で構成されている場合に、ロジカルユニットを列毎に分割管理する情報を登録する。ＬＵ４（２１４）は、管理ＬＵＮに、ＬＵ４－１とＬＵ４－２とし

て登録する関連ＬＵＮは、ＲＡＮＫ１（２０１）に示すように、同一のＲＡＮＫ内に複数のロジカルユニットが設定されている場合に、関連するＬＵＮを登録する。ＬＵ１（２１１）の関連ＬＵＮにはＬＵ２（２１２）を、ＬＵ２（２１２）の関連ＬＵＮにはＬＵ１（２１１）を登録する。

【００３５】かかる情報を登録後、図４に示す電源オフ処理と図５に示す診断処理を起動する。尚、電源オフ処理に代えて、節電モードへ移行させる処理であっても良い。以下、電源オフ処理の場合が節電効果が大きいので、これを中心に実施例を説明するが、電源遮断の代わりに節電モードの選択、実行であっても良い。

【００３６】アクセス時刻と節電開始時刻（電源オフ時刻）は、制御部１６０内の計時機構１６３から現在の時刻を参照して登録する。

【００３７】上位装置１０１からのアクセスを受領したとき、磁気ディスク装置管理テーブル２５０からアクセス対象のロジカルユニットに属する磁気ディスク装置１９０の節電開始時刻を判定し、節電開始時刻が登録されていない場合は、継続してアクセスを実行する。節電開始時刻が登録されている場合は、磁気ディスク装置の電源がオフになっているか又は節電モードにある。このため磁気ディスク装置管理テーブル２５０の磁気ディスク装置位置に登録している磁気ディスク装置１９０の電源を投入し又は節電モードを解除し、節電開始時刻をクリアした後アクセスを実行する。アクセスの実行を終了した後、アクセス時刻を現在の時刻に更新する。

【００３８】図４の電源オフ処理では、設定されているロジカルユニット単位に、磁気ディスク装置管理テーブル２５０のアクセス時刻を監視する。上位装置からアクセスされなくなたロジカルユニットをアクセス時刻に登録された時刻と現在の時刻を比較し、その差が節電待ち時間２６０を超えた時点で、対象のロジカルユニットに属する磁気ディスク装置１９０を、磁気ディスク装置管理テーブル２５０の磁気ディスク装置位置から判断し、磁気ディスク装置位置に登録されている磁気ディスク装置１９０の電源を遮断するか又は節電モードとする。電源を遮断した時刻は、節電開始時刻（図２）に登録する。

【００３９】ＬＵ１（２１１）は、節電待ち時間２６０を超過した場合、登録されている関連ＬＵＮのＬＵ２（２１２）の状態により電源の投入、遮断（節電モードか否か）を判定する。ＬＵ２（２１２）が節電待ち時間を超えていない場合は、磁気ディスク装置管理テーブル２５０の節電開始時刻が登録されおらず、ＬＵ２（２１２）にて磁気ディスク装置１９０を使用していると判定して、その磁気ディスク装置の電源を遮断しない又は節電モードとしない。ＬＵ２（２１２）が節電待ち時間を超えている場合は、磁気ディスク装置管理テーブル２５０の節電開始時刻が登録済みとなり、ＬＵ１（２１１）と共に節電待ち時間２６０をこえていることから、磁気ディスク装置位置の磁気ディスク装置１９０の電源を遮断する。電源を遮断した時刻は、節電開始時刻に登録する。

【００４０】ＬＵＮ４（２１４）は、節電待ち時間２６０を超過した場合、登録されている管理ＬＵＮにて電源のオフを判定する。ＬＵＮ４（２１４）は一つのロジカルユニット構成であるが、複数列の磁気ディスク装置１９０から構成されていることから管理ＬＵＮにて分割して管理する。１列目がＬＵＮ４－１、２列目がＬＵＮ４－２と登録されている。ＬＵＮ４－１は、関連ＬＵＮにＬＵＮ３（２１３）が登録されていることから、ＬＵＮ３（２１３）の節電開始時刻が登録されていない場合は、ＬＵＮ３（２１３）にて使用中であることから１列目の磁気ディスク装置１９０の電源をオフしない。ＬＵＮ４－２は、関連ＬＵＮがないことから、２列目の磁気ディスク装置１９０の電源をオフする。電源をオフした時間は、節電開始時刻に登録する。このようにしてＬＵＮ４では、一つの論理ユニットを構成する磁気ディスク装置群の一部の電源遮断又は節電モードへの移行を可能としている。

【００４１】図５の診断処理では、設定されているロジカルユニット単位に、磁気ディスク装置管理テーブル２５０の節電開始時刻を監視する。節電開始時刻に登録された時刻と現在の時刻を比較し、その差が診断開始時間２７０を超えた時点で、対象のロジカルユニットに属する磁気ディスク装置１９０を、磁気ディスク装置管理テーブル２５０の磁気ディスク装置位置から判断し、磁気ディスク装置位置に登録されている磁気ディスク装置１９０の診断を実行する。診断を終了した後、節電モードを開始する。

【００４２】診断を実行する場合、上位装置１０１からのアクセス処理を優先し、アクセスを受領したときは、診断を停止し、アクセスを実行し、アクセス処理終了後、診断を再開する。

【００４３】磁気ディスク装置１９０の診断は、診断開始時間２７０に２４時間時計の時刻を登録することにより、電源オフしてからの経過時間でなく、登録した時刻になった時点で診断を開始する。また、電源オフしてからの経過時間で診断を開始してもよい。

【００４４】診断を実行する場合、上位装置１０１からのアクセス処理を優先し、アクセスを受領したときは、診断を停止し、アクセスを実行し、アクセス処理終了後、診断を再開する。診断は、磁気ディスク装置管理テーブル２５０に登録されているロジカルユニット単位に同時に実行する。ロジカルユニットが複数列で構成されている場合は、列単位に分けて管理している管理ＬＵＮ単位に実行する。最大列構成の台数の磁気ディスク装置１９０の診断を同時に行う。図１の構成の場合は、図示しない２台の磁気ディスク装置を含めて、各ＲＡＮＫに

おいて５台の磁気ディスク装置の診断を同時に行う。ここで例えば１ＲＡＮＫに４台の磁気ディスク装置が実装されている場合であれば４台の診断を同時に行う。そのＲＡＮＫのディスクアレイとしての論理的磁気ディスク装置を使用する前にチェックするためである。

【００４５】尚、特殊な態様ではあるが、１ＲＡＮＫに１台のみの磁気ディスク装置を有するディスクアレイも、本発明を適用できる。この場合には１組の磁気ディスク装置群は１台の磁気ディスク装置で構成されることとなる。

【００４６】スペア用磁気ディスク装置１９１の診断は、一定時間経過した時点で行う。診断開始時間２７０に２４時間時計の時刻が登録されている場合は、登録されている時刻になった時点で行ってもよい。

【００４７】節電開始時刻２６０と診断開始時間は、ホストコマンド、操作パネル１３０、ＬＡＮポート１４０およびＲＳ２３２Ｃポートから変更できる。変更の指示を受けたときにメモリ１６２の情報を更新する。ホストコマンドによる変更は、これらの情報の変更を指示するベンダーユニークなコマンドによって行う。操作パネル１３０による変更は、操作パネル上にこれらの情報の変更設定メニューを表示し、オペレータが変更したい値を入力することにより行う。ＬＡＮポートおよびＲＳ２３２Ｃポートによる変更は、ＬＡＮおよびＲＳ２３２Ｃに特定のインタフェースを規定し、インタフェース内に変更する情報を埋め込み、ＬＡＮおよびＲＳ２３２Ｃの接続先の装置から情報を送ることにより行う。変更された情報は、変更の指示があった時点で、メモリ１６２に書き込み、書き込み終了により有効となる。

【００４８】

【発明の効果】本発明は、ディスクアレイ装置に実装されている磁気ディスク装置の消費電力を抑えることができ、さらに、上位装置からアクセスされない磁気ディスク装置を対象とすることから、ディスクアレイ装置の著しい性能低下を抑制しつつ省エネルギーを可能とする。また、電源を遮断した磁気ディスク装置に対して診断を行うことにより、しばしば電源を遮断することとなる磁気ディスク装置の信頼性を確認することができる。

【００４９】より具体的には、例えば、回転起動時に２０Ｗの電力が必要な１台の磁気ディスク装置において、その電子回路ではリードライト時に５．５Ｗ、リードライトアイドル時に３．５Ｗの電力消費があり、その磁気ディスク装置のスピンドルがアイドル回転時には４Ｗから５Ｗの電力消費がある。この場合において、電子回路の機能を停止する節電モードを選択すれば、少なくとも３．５Ｗの電力消費が抑制され、スピンドルの回転も止めれば、これに加えて４Ｗから５Ｗの電力消費が抑制されることとなる。

【００５０】実際にはＲＡＮＫ単位で節電するため、５列構成のディスクアレイであれば、１つのＲＡＮＫの電子回路が節電モードとなっていれば、約１８Ｗの電力消費が、スピンドルの回転も止めればこれに加えて２０Ｗから２５Ｗの電力消費が抑制されることとなる。ディスクアレイ装置の使い方によりこれらの数値は増減がある。例えば、２４時間連続運転における昼夜のアクセス需要の変化、これに伴う節電モードの実行時間によって、節電効果は大きくなる。

【図面の簡単な説明】

【図１】本発明の１実施例に係わるディスクアレイ装置の全体構成図である。

【図２】本発明の１実施例に係わる管理テーブルの図である。

【図３】本発明の１実施例に係わる磁気ディスク装置電源オン動作フローチャートである。

【図４】本発明の１実施例に係わる磁気ディスク装置電源オフ動作フローチャートである。

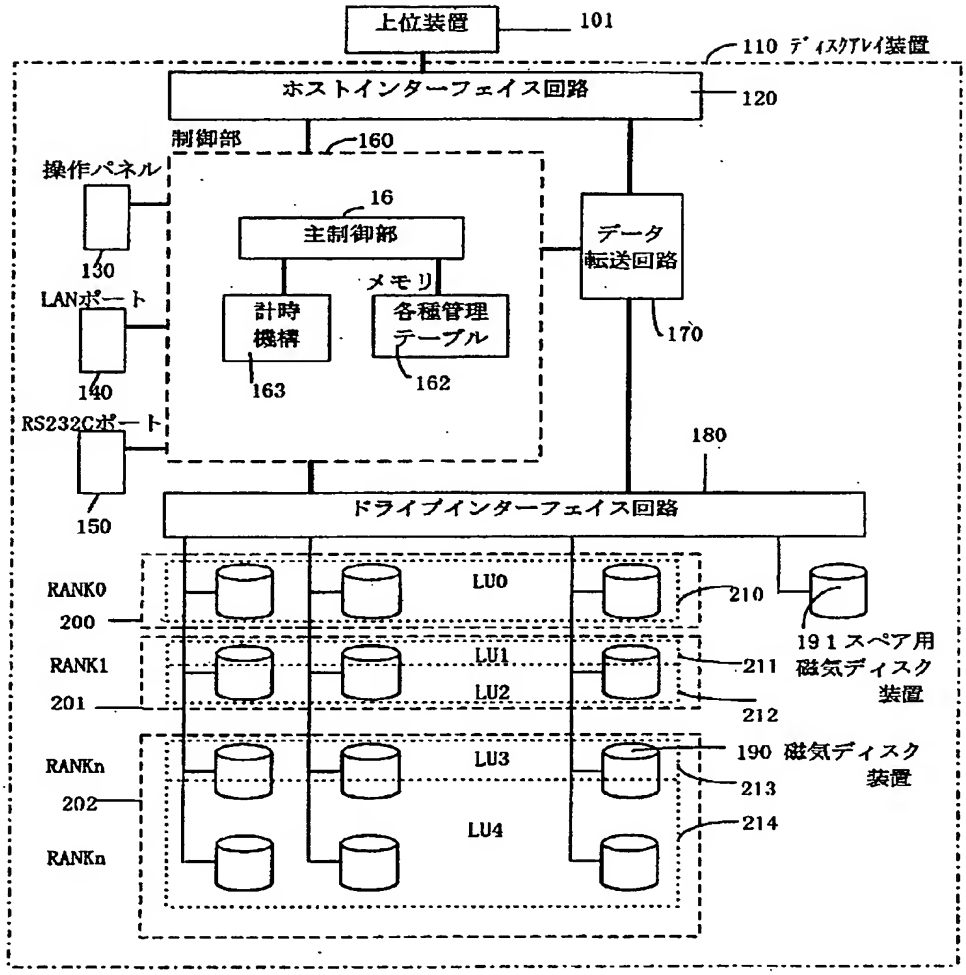
【図５】本発明の１実施例に係わる磁気ディスク装置電源診断動作フローチャートである。

【符号の説明】

１０１…上位装置、　１１０…ディスクアレイ装置、１２０…ホストインタフェース回路、１３０…操作パネル、１４０…ＬＡＮ（ローカルエリアネットワーク）ポート、１５０…ＲＳ２３２Ｃポート、１６０…制御部、１６１…主制御部、１６２…メモリ、１６３…計時機構、１７０…データ転送回路、１８０…ドライブインタフェース回路、１９０…磁気ディスク装置、１９１…スペア用磁気ディスク装置、２００…ＲＡＮＫ０、２０１…ＲＡＮＫ１、　２０２…ＲＡＮＫ２、２１０…ＬＵ（ロジカルユニット）０、２１１…ＬＵ（ロジカルユニット）１、２１２…ＬＵ（ロジカルユニット）２、２１３…ＬＵ（ロジカルユニット）３、２１４…ＬＵ（ロジカルユニット）４、２５０…磁気ディスク装置管理テーブル、２６０…節電待ち時刻２７０…診断開始時間。

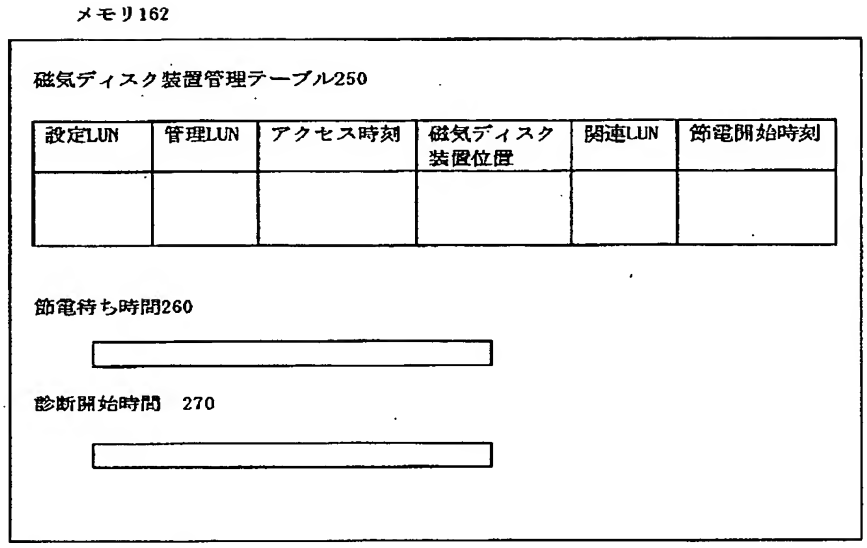
【図１】

図１


上位装置
101
110 ディスクアレイ装置
ホストインターフェイス回路
120
制御部
160
操作パネル
130
LANポート
140
RS232Cポート
150
16
主制御部
メモリ
計時機構
163
各種管理テーブル
162
データ転送回路
170
180
ドライブインターフェイス回路
RANK0
200
RANK1
201
RANKn
202
RANKn
LU0
LU1
LU2
LU3
LU4
210
211
212
213
214
191 スペア用磁気ディスク装置
190 磁気ディスク装置

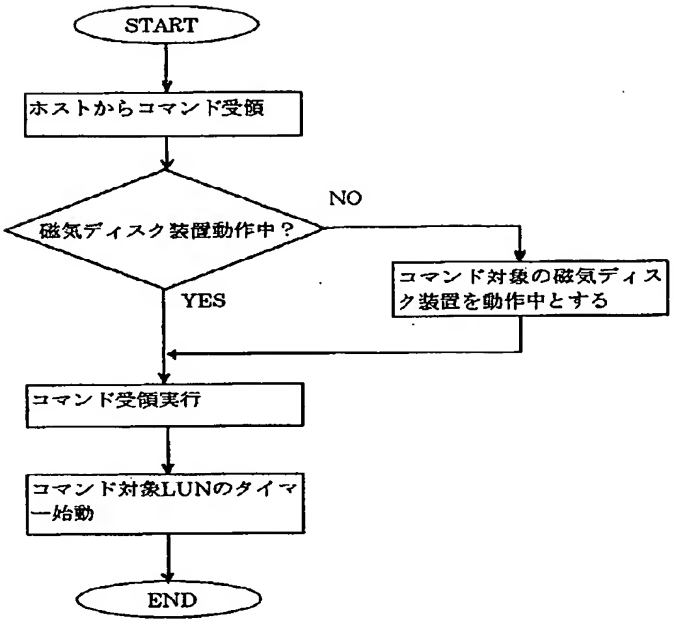
【図２】

図２


メモリ162
磁気ディスク装置管理テーブル250
設定LUN
管理LUN
アクセス時刻
磁気ディスク装置位置
関連LUN
節電開始時刻
節電待ち時間260
診断開始時間　270

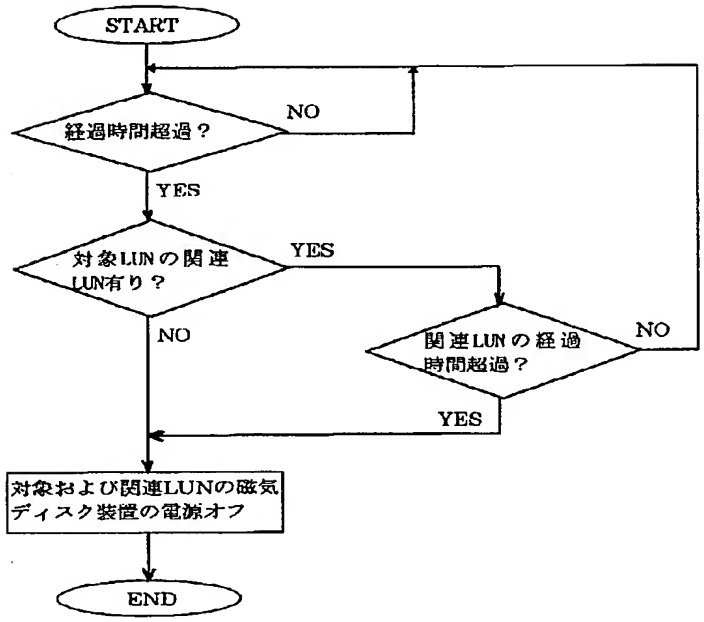
【図３】

図３


START
ホストからコマンド受領
磁気ディスク装置動作中？
NO
YES
コマンド対象の磁気ディスク装置を動作中とする
コマンド受領実行
コマンド対象LUNのタイマー始動
END

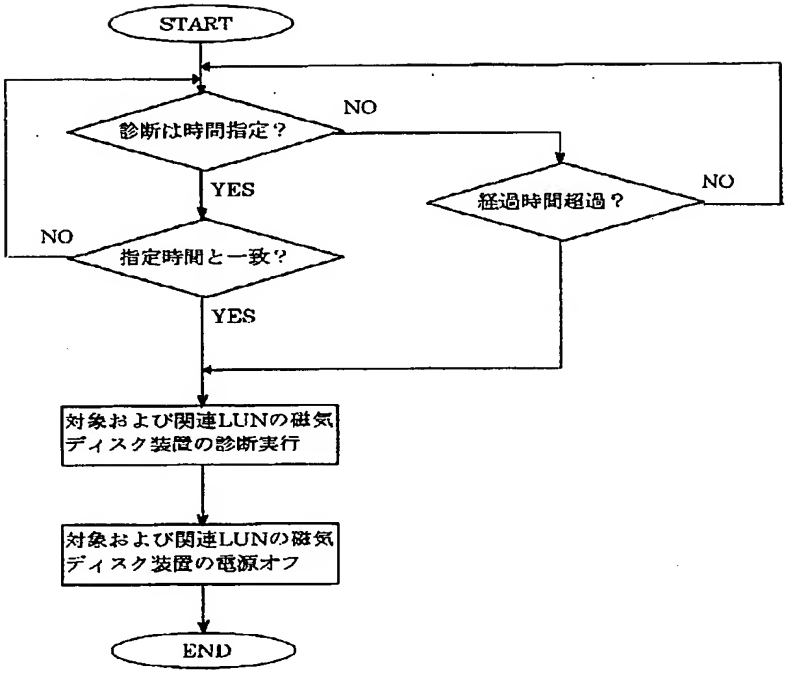
【図４】

図４


START
経過時間超過？
NO
YES
対象LUNの関連LUN有り？
YES
NO
関連LUNの経過時間超過？
NO
YES
対象および関連LUNの磁気ディスク装置の電源オフ
END

【図５】

図５



フロントページの続き

(72)発明者　小林　正明
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

35　Ｆターム(参考)　5B011 EB07 LL14
5B065 BA01 CA16 CA30 CC01 ZA14
5D066 BA02 BA05